Robust Tensor Factorisations

New Algorithms and Extensive Comparisons

Mehdi Bahri

MSc Advanced Computing
Imperial College London

Septembre 2016

Outline

Outline

Motivations

- **Tensor representations**: preserve structural information (correlation)
- **Low-rank modelling**: low-rank data assumption, data compression
- **Robustness**: Noisy or corrupt data
- **Context**: Compressed Sensing, inverse problems, signal processing

Compressed Sensing

Fixed-rate sampling: Nyquist (1928), Shannon (1949). **Sufficient conditions, *not* necessary**.

Linear inverse problem:

$$\mathcal{A}\mathbf{x} = \mathbf{y}$$

$\mathcal{A}$ *sensing matrix*. In general: underdetermined system. Numerically difficult.

Exact reconstruction is possible if:

- Signals are *sparse* in some domain: **redundant information**
- Signals are *incoherent*: **largest correlation is small**

Compressed Sensing

Restricted Isometry Property (RIP) (Candes & Tao 2005): $\mathcal{A}$ sufficiently close to an isometry. Let $\mathbf{x}$ s-sparse:

$$(1-\delta_s)||\mathbf{x}||_{\ell_2}^2 \leq ||\mathcal{A}\mathbf{x}||_{\ell_2}^2 \leq (1+\delta_s)||\mathbf{x}||_{\ell_2}^2$$

Sufficient condition of incoherence for sparse vectors.

Robust PCA (Candes et. al. 2011)

$$\min_{\mathbf{A},\mathbf{E}} \ \text{rank}(\mathbf{A}) + ||\mathbf{E}||_0 \quad \text{s.t} \quad \mathbf{X} = \mathbf{A} + \mathbf{E}$$

NP-Hard. Convex relaxation (*Basis Pursuit Problem*):

$$\min_{\mathbf{A},\mathbf{E}} \ ||\mathbf{A}||_* + ||\mathbf{E}||_1 \quad \text{s.t} \quad \mathbf{X} = \mathbf{A} + \mathbf{E}$$

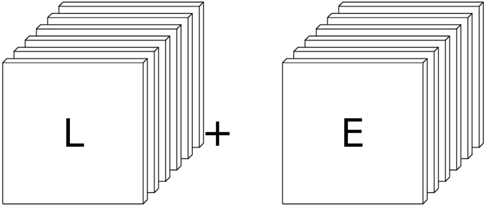
Exact solution by optimisation with overwhelming probability (≈ 1).

Outline

Non-orthogonal Robust 2D PCA

Model

$\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ ordre 3. Samples $\leftrightarrow$ frontal slices.
Observation model: $\mathbf{X}_n = \mathbf{U_c} \mathbf{T}_n \mathbf{U_r}^{\mathsf{T}} + \mathbf{E}_n$.



Robust instance of *Tucker2* without orthogonality

$$\mathcal{X} = \mathcal{L} + \mathcal{E} \quad \text{s.t} \quad \mathcal{L} = \mathcal{T} \times_1 \mathbf{U_c} \times_2 \mathbf{U_r}$$

$$\mathcal{X} \in \mathbb{R}^{n \times m \times N} \quad \mathbf{U_c} \in \mathbb{R}^{n \times r} \quad \mathbf{U_r} \in \mathbb{R}^{m \times r} \quad \mathcal{T} \in \mathbb{R}^{r \times r \times N}$$

Non-orthogonal Robust 2D PCA

Model

Optimisation problem:

$$\begin{array}{ll} \min_{\mathbf{U_c},\mathbf{U_r},\mathcal{T},\mathcal{E}} & f(\mathbf{U_c},\mathbf{U_r},\mathcal{T},\mathcal{E}) \\ \text{s.t} & \forall n,\ \mathbf{X}_n = \mathbf{U_c}\mathbf{T}_n{\mathbf{U_r}}^\mathsf{T} + \mathbf{E}_n \end{array}$$

Restricted form (f):

$$\alpha_c f(\mathbf{U_c}) + \alpha_r f(\mathbf{U_r}) + \alpha_t \sum_n g(\mathbf{T}_n) + \lambda\Omega(\mathcal{E})$$

- Separable penalties, different choices possible
- Non-linear, non-convex, constrained problem
- Ω non-smooth

Non-orthogonal Robust 2D PCA

Base case

$$f(\ldots) = \frac{\alpha_c}{2}||\mathbf{U_c}||_F^2 + \frac{\alpha_r}{2}||\mathbf{U_r}||_F^2 + \frac{\alpha_t}{2}\sum_n ||\mathbf{T}_n||_F^2 + \lambda \sum_n ||\mathbf{E}_n||_1$$

- Schatten-2 (Frobenius) norm on the bases: low rank
- ℓ_1 norm on $\mathbf{E}_n$, eq. to $||\boldsymbol{\mathcal{E}}||_1$: unstructured noise
- Frobenius norm on $\mathbf{T}_n$: Tikhonov regularisation for numerical stability

Non-orthogonal Robust 2D PCA

Numerical solution

Alternating Directions Method (ADMM)

- Primal-dual algorithm
- Based on the *Augmented Lagrangian*
- Solve for each direction independently in turn
- Stopping criterion: first order conditions (primal feasibility)

Sub-problems:

- *penalized least squares*, convex
- FONC if smooth
- Non-smooth component ($||.||_1$): **proximal mapping**
- $\mathbf{T}_n$: Stein equation $\mathbf{X} + \mathbf{AXB} = \mathbf{C} \rightarrow$ Hessenberg-Schur

Non-orthogonal Robust 2D PCA

Sparse core

$$f(\ldots) = \frac{\alpha_c}{2}||\mathbf{U_c}||_\mathrm{F}^2 + \frac{\alpha_r}{2}||\mathbf{U_r}||_\mathrm{F}^2 + \alpha_t \sum_n ||\mathbf{T}_n||_1 + \lambda \sum_n ||\mathbf{E}_n||_1$$

Inspiration: *sparse dictionary learning*.

$\ell_2 \to \ell_1$: automatically discard redundant elements.

Numerical solution by splitting

Introducing $\mathcal{K}$ s.t $\forall n,\ \mathbf{K}_n = \mathbf{T}_n$, augmented problem:

$$\begin{array}{ll} \min_{\mathbf{U_c},\mathbf{U_r},\mathcal{T},\mathcal{E}} & f(\mathbf{U_c},\mathbf{U_r},\mathcal{T},\mathcal{E}) \\ \text{s.t} & \forall n,\ \mathbf{X}_n = \mathbf{U_c}\mathbf{K}_n{\mathbf{U_r}}^\mathrm{T} + \mathbf{E}_n \\ & \forall n,\ \mathbf{T}_n = \mathbf{K}_n \end{array}$$

Non-orthogonal Robust 2D PCA

Sparse core

Why ?
Augmented Lagrangian:

$$\mathcal{L}(\dots) = \alpha_t \sum_n ||\mathbf{T}_n||_1 + \frac{\mu}{2} \sum_n ||\mathbf{X}_n - \mathbf{U_c} \mathbf{T}_n \mathbf{U_r}^\mathsf{T} - \mathbf{E}_n||_\mathsf{F}^2 + \dots$$

Augmented Lagrangian with splitting:

$$\mathcal{L}(\dots) = \alpha_t \sum_n ||\mathbf{T}_n||_1 + \frac{\mu \mathcal{T}}{2} \sum_n ||\mathbf{T}_n - \mathbf{K}_n||_\mathsf{F}^2 + \frac{\mu}{2} \sum_n ||\dots||_\mathsf{F}^2 + \dots$$

$\Rightarrow$ proximal operator of $||.||_1$.

Alternative ?
Linearisation. Challenge: numerically unstable.

Non-orthogonal Robust 2D PCA

Group Lasso variant

Group Lasso: ℓ_1/ℓ_2 norm. Group: columns.

$$f(\ldots) = \alpha_c||\mathbf{U_c}||_{\ell_1/\ell_2} + \alpha_r||\mathbf{U_r}||_{\ell_1/\ell_2} + \alpha_t \sum_n h(\mathbf{T}_n) + \lambda \sum_n ||\mathbf{E}_n||_1$$

$$h = ||.||_F^2 \text{ or } ||.||_1$$

Penalises $\dim(\mathrm{Span}(\mathbf{U_c}))$ and $\dim(\mathrm{Span}(\mathbf{U_r}))$. *Structural* approach as opposed to *spectral*.

Numerical solution: ADMM with splitting.

Non-orthogonal Robust 2D PCA

Robust estimator of the mean

Iterative procedure.

$$\mathsf{M}^t = \frac{1}{N}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{E}}^{t+1})$$

Automated removal of the outliers. Selective trimmed mean.

Empirically:

- Improved or comparable performance (*background subtraction*, some gross corruptions)
- Converges to the sample mean of $\boldsymbol{\mathcal{L}}$

Non-orthogonal Robust 2D PCA

Summary

Components:

- Pair of dictionaries $\mathbf{U_c}$, $\mathbf{U_r}$
- *Core* tensor $\boldsymbol{\mathcal{T}}$: bound $r \geq$ *true rank*
- *Outliers* tensor $\boldsymbol{\mathcal{E}}$: unstructured noise

Rank minimisation of $\mathbf{U_c T}_n \mathbf{U_r}$:

- *Spectral* approach: Schatten-2 norm
- *Structural* approach: ℓ_1/ℓ_2 mixed-norm

Regularisation of $\boldsymbol{\mathcal{T}}$: either *dense* ℓ_2 or *sparse* ℓ_1

Abbreviations:

Core vs Bases	ℓ_2	ℓ_1/ℓ_2
ℓ_2	RPCA2D Fro ℓ_2	RPCA2D GL ℓ_2
ℓ_1	RPCA2D Fro ℓ_1	RPCA2D GL ℓ_1

Non-orthogonal Robust 2D PCA

Implementation and Complexity Analysis

Implementation:

- Matlab 2015b + toolboxes
- MMX for fast tensor-matrix products, C++, OpenMP
- Personal MEX extensions, C, BLAS/LAPACK, OpenMP

$O(N(mnr + (m+n)r + mn + \min(m,n)r^2 + r^3 + r^2))$ FLOP/it:

- Linear in N
- Homogeneous in m, n, depends on min: best for rectangular matrices *tall & skinny*, *short & fat* ...
- Cubical in r

$O(N(mn + (m+n)r + r^2))$ or $O(Nmn + (m+n)r + r^2)$ space

Non-orthogonal Robust 2D PCA

Experimental validation

100 matrices, 1000×1000. 30 and 60% noise.
True rank $\mathbf{U_c}$: 23, $\mathbf{U_r}$: 41.

Parameters:

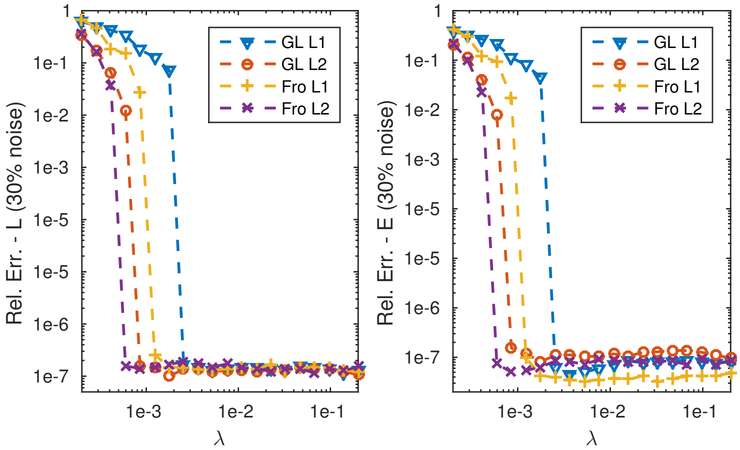
- λ varies
- $\alpha_c = \alpha_r = 1$ if Fro, $1e - 3$ if GL
- $\alpha_t = 1e - 2$
- Max rank $= 100$

Measures: relative ℓ_2-norm error, *nnz* $\boldsymbol{\mathcal{E}}$, mean FSIM $\boldsymbol{\mathcal{L}}$.

Non-orthogonal Robust 2D PCA

Experimental validation

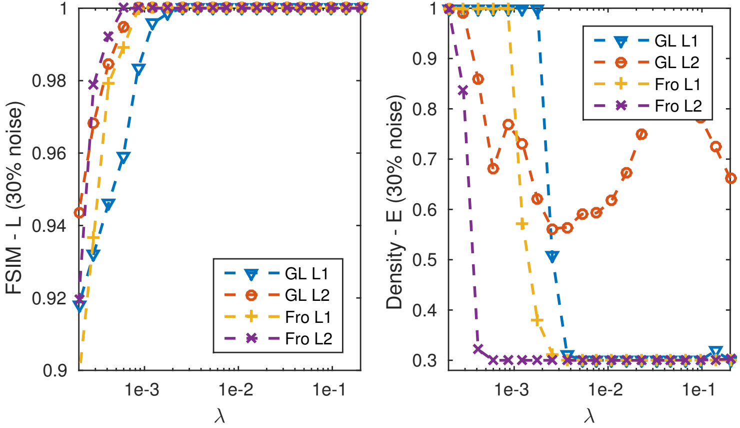
Exact reconstruction:



Non-orthogonal Robust 2D PCA

Experimental validation

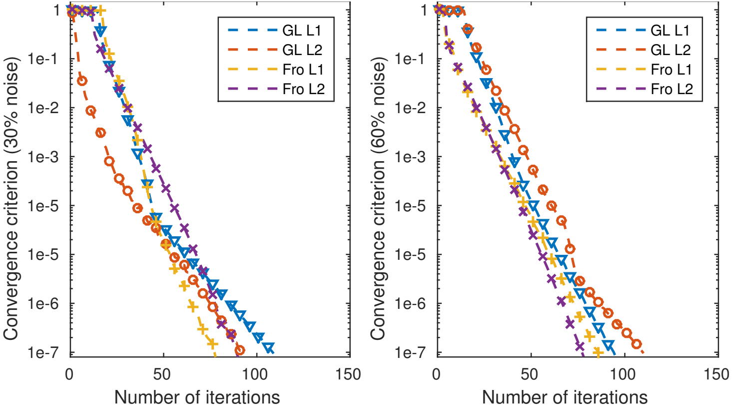
Exact structural recovery (except *density* RPCA2D GL ℓ_2)



Non-orthogonal Robust 2D PCA

Experimental validation

Linear convergence. **Important:** non-convex problem, initialisation matters.



Non-orthogonal Robust 2D PCA

New view on the problem

SDL: $\min \sum_n ||\mathbf{D}\mathbf{r}_n - \mathbf{x}_n||_2^2 + \lambda||\mathbf{r}_n||_1$
Recall the observation model:

$$\mathbf{X}_n = \mathbf{U_c}\mathbf{T}_n\mathbf{U_r}^T + \mathbf{E}_n$$

Non-orthogonal Robust 2D PCA

New view on the problem

SDL: $\min \sum_n ||\mathbf{D}\mathbf{r}_n - \mathbf{x}_n||_2^2 + \lambda||\mathbf{r}_n||_1$
Recall the observation model:

$$\mathbf{X}_n = \mathbf{U_c}\mathbf{T}_n{\mathbf{U_r}}^T + \mathbf{E}_n \rightarrow \text{vec}(\mathbf{X}_n) = (\mathbf{U_r} \otimes \mathbf{U_c})\text{vec}(\mathbf{T}_n) + \text{vec}(\mathbf{E}_n)$$

Note:

$$||\mathbf{A}||_1 = ||\text{vec}(\mathbf{A})||_1$$

Non-orthogonal Robust 2D PCA

New view on the problem

AL of RPCA2D Fro ℓ_1:

$$\frac{\mu}{2}\sum_n ||\mathbf{U_c}\mathbf{T}_n\mathbf{U_r}^\mathsf{T} + \mathbf{E}_n - \mathbf{X}_n||_\mathsf{F}^2 + \alpha_t \sum_n ||\mathbf{T}_n||_1 + \lambda \sum_n ||\mathbf{E}_n||_1 + \frac{\alpha_c}{2}||\mathbf{U_c}||_\mathsf{F}^2 + \frac{\alpha_r}{2}||\mathbf{U_r}||_\mathsf{F}^2 + \dots$$

Non-orthogonal Robust 2D PCA

New view on the problem

AL of RPCA2D Fro ℓ_1:

$$\frac{\mu}{2}\sum_n ||\mathbf{U_c}\mathbf{T}_n\mathbf{U_r}^{\mathsf{T}} + \mathbf{E}_n - \mathbf{X}_n||_{\mathsf{F}}^2 + \alpha_t \sum_n ||\mathbf{T}_n||_1 + \lambda \sum_n ||\mathbf{E}_n||_1 + \frac{\alpha_c}{2}||\mathbf{U_c}||_{\mathsf{F}}^2 + \frac{\alpha_r}{2}||\mathbf{U_r}||_{\mathsf{F}}^2 + \dots$$

Vectorising:

$$\frac{\mu}{2}\sum_n ||(\mathbf{U_r}\otimes\mathbf{U_c})\mathrm{vec}(\mathbf{T}_n) + \mathrm{vec}(\mathbf{E}_n) - \mathrm{vec}(\mathbf{X}_n)||_{\mathsf{F}}^2 + \alpha_t \sum_n ||\mathrm{vec}(\mathbf{T}_n)||_1 + \lambda \sum_n ||\mathrm{vec}(\mathbf{E}_n)||_1 + \frac{\alpha_c}{2}||\mathbf{U_c}||_{\mathsf{F}}^2 + \frac{\alpha_r}{2}||\mathbf{U_r}||_{\mathsf{F}}^2 + \dots$$

Non-orthogonal Robust 2D PCA

A theorem on Schatten norms

Theorem

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{B} \in \mathbb{R}^{p \times q}$, *then:*

$$\forall p > 0, ||\mathbf{A} \otimes \mathbf{B}||_p = ||\mathbf{A}||_p ||\mathbf{B}||_p$$

Where $||.||_p$ *denotes the Schatten-p norm.*

Note: Also true for (mixed) element-wise norms by construction.

Non-orthogonal Robust 2D PCA

A theorem on Schatten norms

Proof.

Let the SVDs $\mathbf{A} = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^T,\ \mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T$ then:

$$\mathbf{A} \otimes \mathbf{B} = (\mathbf{U}_A \otimes \mathbf{U}_B)(\mathbf{\Sigma}_A \otimes \mathbf{\Sigma}_B)(\mathbf{V}_A^T \otimes \mathbf{V}_B^T)$$

so the singular values of $\mathbf{A} \otimes \mathbf{B}$ are the $\sigma_{A,i}\sigma_{B,j}$ and:

$$\begin{aligned} ||\mathbf{A} \otimes \mathbf{B}||_p &= (\sum_{i,j} (\sigma_{A,i}\sigma_{B,j})^p)^{1/p} \\ &= (\sum_i \sigma_{A,i}^p)^{1/p} (\sum_j \sigma_{B,j}^p)^{1/p} \\ &= ||\mathbf{A}||_p ||\mathbf{B}||_p \end{aligned}$$

$\square$

Non-orthogonal Robust 2D PCA

Putting everything together

Therefore:

$$||\mathbf{U_r} \otimes \mathbf{U_c}||_F = ||\mathbf{U_r}||_F ||\mathbf{U_c}||_F \leq \frac{1}{2}(||\mathbf{U_r}||_F^2 + ||\mathbf{U_c}||_F^2)$$

Variational charact. of Nuclear norm:

$$||\mathbf{U_c}\mathbf{U_r}^T||_* = \inf \frac{1}{2}(||\mathbf{U_c}||_F^2 + ||\mathbf{U_r}||_F^2)$$

Hölder's inequality:

$$||\mathbf{AB}||_* \leq ||\mathbf{A}||_F ||\mathbf{B}||_F$$

Non-orthogonal Robust 2D PCA

Putting everything together

- Minimise $||\mathbf{U_r} \otimes \mathbf{U_c}||_F$
- Minimise $||\mathbf{U_c}\mathbf{U_r}^T||_*$
- Structured (Kronecker-decomposable) dictionary
- Robust model

Outline

Background subtraction

Presentation

Video

- *Highway* - static background
- *Airport Hall* - changing background
- *Ground-truth* available

Methodology

- Optimise no more than 2 params / method
- *grid search*
- Robust mean estimator enabled
- Performance metric: AUC

Background subtraction

Results

Five best performances (*Highway*)

BRTF	TRPCA16	RPCA2D GL ℓ_1	RPCA2D Fro ℓ_1	RCPD
0.9451	0.9449	0.9432	0.943	0.936

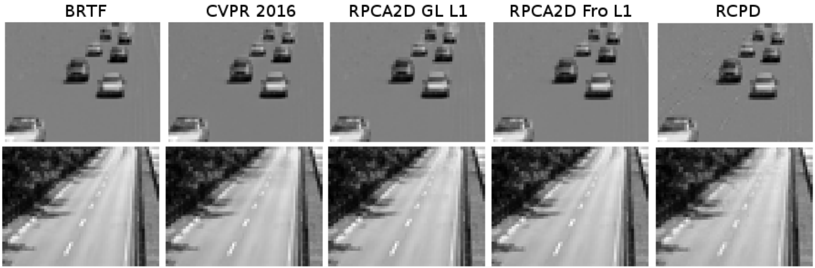
Five best performances (*Airport Hall*)

RPCA2D GL ℓ_1	RPCA2D ℓ_1	RPCA2D ℓ_2	TRPCA16	NCTRPCA
0.895	0.884	0.883	0.864	0.863

Background subtraction

Results

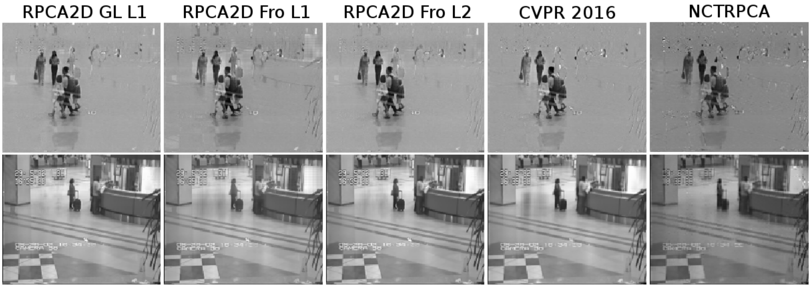
Five best performances (*Highway*)



Background subtraction

Results

Five best performances (*Airport Hall*)

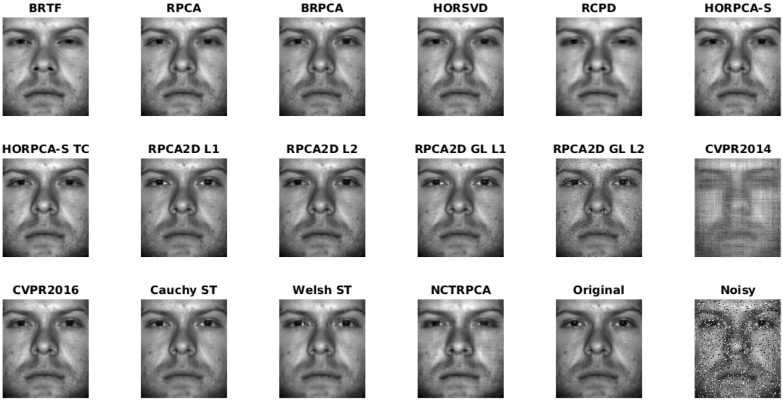


Denoising of grayscale face images

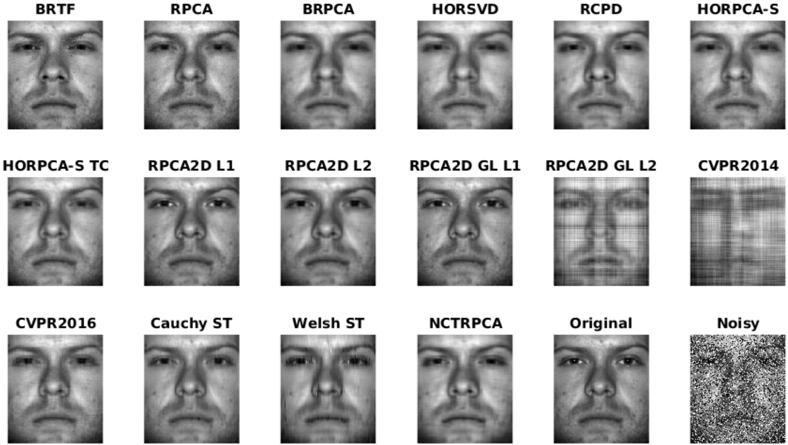
Five best results by FSIM

	1	2	3	4	5
10%	2D GL L1 0.987	2D L1 0.983	Cauchy ST 0.983	2D L2 0.979	TRPCA16 0.978
30%	2D L1 0.957	2D L2 0.947	2D GL L1 0.947	Cauchy ST 0.942	HORPCA-S TC 0.935
60%	2D L1 0.896	RCPD 0.882	2D L2 0.865	NCTRPCA 0.851	TRPCA16 0.843

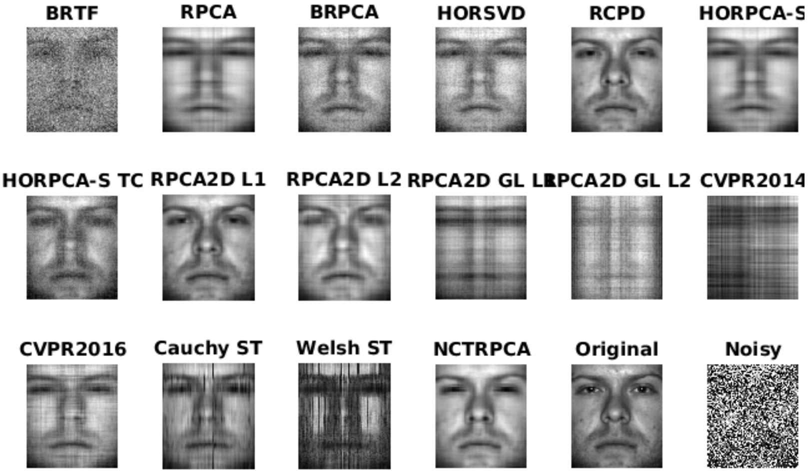
10% noise


BRTF
RPCA
BRPCA
HORSVD
RCPD
HORPCA-S
HORPCA-S TC
RPCA2D L1
RPCA2D L2
RPCA2D GL L1
RPCA2D GL L2
CVPR2014
CVPR2016
Cauchy ST
Welsh ST
NCTRPCA
Original
Noisy

30% noise


BRTF
RPCA
BRPCA
HORSVD
RCPD
HORPCA-S
HORPCA-S TC
RPCA2D L1
RPCA2D L2
RPCA2D GL L1
RPCA2D GL L2
CVPR2014
CVPR2016
Cauchy ST
Welsh ST
NCTRPCA
Original
Noisy

60% noise

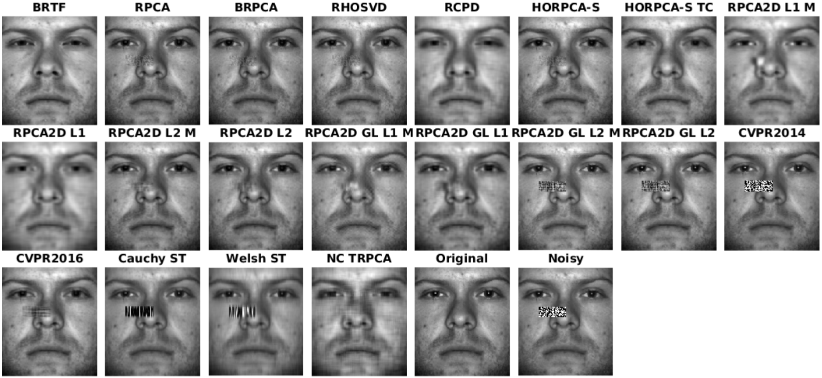

BRTF
RPCA
BRPCA
HORSVD
RCPD
HORPCA-S
HORPCA-S TC
RPCA2D L1
RPCA2D L2
RPCA2D GL L1
RPCA2D GL L2
CVPR2014
CVPR2016
Cauchy ST
Welsh ST
NCTRPCA
Original
Noisy

Partial obstruction of grayscale face images

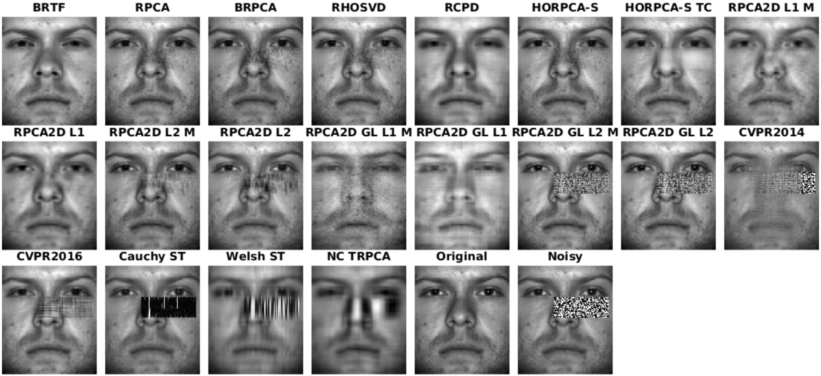
Five best performances by PSNR

	1	2	3	4	5
50	BRPCA 38.6465	RPCA 37.585	HORPCA-S 37.3662	RHOSVD 37.264	2D L2 33.7598
100	BRPCA 33.9074	RPCA 32.7486	HORPCA-S 32.5031	RHOSVD 32.0228	2D L2 29.8869
160	BRPCA 29.4723	RPCA 28.2569	HORPCA-S 28.0179	RHOSVD 27.6087	2D L2 25.413

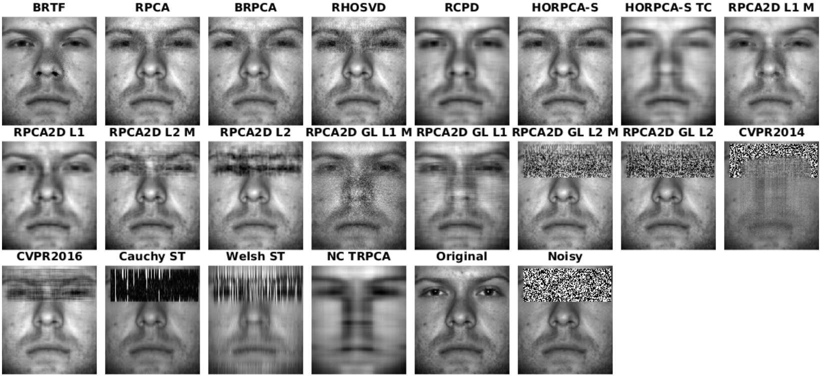
Max size of 50



Max size of 100



Max size of 160



Outline

Bayesian Model

Overview

Motivations

- Success of *sparse Bayesian learning*
- No hyperparameter optimisation
- Few existing Bayesian models for *robust low-rank modelling*
- No Bayesian treatment of the *Tucker* decomposition

Observation model

$$\mathcal{X} = \mathcal{T} \times_1 \mathbf{U_c} \times_2 \mathbf{U_r} + \mathcal{E} + \mathcal{N}$$

$\mathcal{N}$: Gaussian white noise

Hypotheses

Independence of the frontal slices:

$$p(\mathcal{X}) = \prod_n p(\mathbf{X}_n) \quad p(\mathcal{T}) = \prod_n p(\mathbf{T}_n)$$

Outliers i.i.d:

$$p(\mathcal{E}) = \prod_{i,j,n} p(e^n_{i,j})$$

Bases:

$$p(\mathbf{U_c}) = \prod_i p(\mathbf{u}_{ci}) \quad p(\mathbf{U_r}) = \prod_i p(\mathbf{u}_{ri})$$

Regularisation and priors

$\ell_2 \rightarrow$ Normal distribution $\ell_1 \rightarrow$ Laplace distribution

Problem: Intractable integrals with Laplace distributions

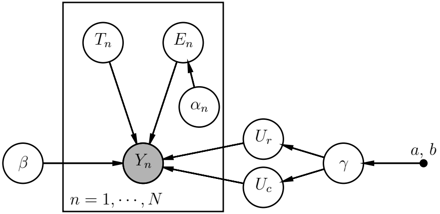
Solution: Hierarchical model: Std Normal + prior on the precision

Inference

Priors

- Precisions: Jeffreys (the *sparsest*)
- Other variables: Normal distributions (*c.f.* Thesis)

Graphical model



Approximate inference: Variational Bayes (mean-field)

Outline

Conclusion

Assessment of the results:

- Good performance on a range of CV applications
- Robust factorisation for third-order tensors
- Possible applications outside of CV (data compression...)
- *Sparse* methods more robust overall

Future work and extensions:

- Publication in top conferences and journals (work in progress)
- Extensions to other norms, missing values
- Sparse and/or scalable implementation